Presupposition Projection Theories Through The Lens of English and Mandarin Large Language Models

Anonymous ACL submission

Abstract

Presupposition projection remains a critical area of linguistic research, particularly in understanding how meaning is inferred beyond explicit assertion. This study explores the processing of presuppositions in conditional sentences by large language models (LLMs) in both English and Mandarin, evaluating their alignment with established linguistic theories such as Satisfaction Theory (ST) and Discourse Representation Theory (DRT). Through controlled experiments inspired by Romoli's (2011) human subject study, we reveal considerable variation across models, both within and across languages, challenging the assumption that LLMs uniformly approximate human-like pragmatic competence. While some models exhibited patterns aligning with ST, others diverged significantly, suggesting that LLMs can produce contextually appropriate text without a structured, human-like understanding of presupposition.

1 Introduction

While much research probed syntactic, semantic, and more recently pragmatic knowledge of LLMs (Hong et al., 2024) (Sieker and Zarrieß, 2023), there has been little examination of the fit between LLMs' behaviour and specific linguistic theories, and even less cross-linguistic work in this domain. However, linguistic theories can provide a valuable source of insight for the goal of explainability in LLMs (Zhao et al., 2024).

This paper considers how modern English and Mandarin LLMs process presuppositions, comparing their behavior to a human baseline and exploring fit with two major theories drawn from the linguistic literature. Strikingly, language models both within and across languages vary considerably in their fit with theoretical models and in their approximation to human behavior in this domain. This fact suggests that LLMs can generate convincingly human-like text while lacking the human-like understanding of pragmatic elements, such as presupposition.

2 Related Work

2.1 Presuppositions

Presuppositions are assumptions that must hold for an utterance to make sense (Beaver et al., 2024). For example, "I turned in my dissertation" presupposes the existence of a dissertation, that it is the speaker's, and so forth. These assumptions can be triggered by specific words (e.g., "my") or arise pragmatically (e.g., that the speaker speaks English).

A key property of presuppositions is projection: they often survive under negation and conditionals. For instance, "I completed my essay in time" presupposes an essay exists; this remains even in "I didn't complete my essay in time" or "If I completed my essay in time…" Linguists have offered two main theoretical accounts to model this phenomenon: Satisfaction Theory (ST) (Heim, 2002) and Discourse Representation Theory (DRT) (Kamp, 1981) (Geurts, 1996). Both offer broad computational-level accounts of semantic and pragmatic interpretation, differing in a number of respects. For our purposes, the crucial differences involve predictions about conditionals.

(1) If Jack killed the man, the weapon he used is hard to find.

(2) If Jack killed the man, his friend was involved.

ST predicts conditional presuppositions here as a default: "If Jack killed the man, he (used a weapon/has a friend)". The former seems correct, while the latter seems too weak (Geurts, 1996). By contrast, DRT predicts unconditional presuppositions for both ("Jack (used a weapon/has a friend)"), with the mirror-image problem. (1) does not intuitively presuppose "Jack used a weapon" as DRT predicts as a default.

ST theorists have claimed that the strengthened presupposition of (2) is the result of additional pragmatic reasoning, due to the lack of a causal or inferential connection between "Jack killed the man" and "Jack has a friend". Romoli (2011) compare the competing theories in a human subjects experiment, manipulating the presence of a causal connection, with results support ST for human pragmatic processing.

2.2 Large Language Models Probing

Recent LLM research has shifted from syntactic and semantic probing to pragmatics, particularly implicatures and presuppositions. Datasets such as ImpPress (Jeretic et al., 2020), ProPress (Asami and Sugawara, 2023), and NOPE (Parrish et al., 2021) assess presupposition understanding, with PUB (Sravanthi et al., 2024) integrating these into broader benchmarks. Findings suggest advanced LLMs increasingly mirror human intuition.

However, one key problem still remains: do LLMs exhibit genuine linguistic structure or replicate training data? Blevins et al. (Blevins et al., 2023) showed structured prompting enables abstraction beyond memorization. Studies on Maximize Presupposition! (Sieker and Zarrieß, 2023) and causal inference (Hong et al., 2024) highlight model variability. Both shows that there is a type of structure in LLMs – they are not merely replicating what they have seen.

LLMs also inform linguistic theory. Cho et al. (Cho and Kim, 2024) found GPT-2 and BERT favor pragmatic scalar implicatures, with GPT-2 relying more on context. Tsvilodub et al. (Tsvilodub et al., 2024) replicated human studies on disjunctions, aligning LLM results with human data. This paper examines LLM processing of presuppositions in consequents, crucial for human-like discourse. Following the practice of Tsvilodub, this paper will replicate the Romoli paper mentioned in section 2.1. Many benchmarks assess presuppositions, making it a key area in NLP advancement.

3 Methodology

3.1 Overview of Romoli's Experiments' Methodology

The previous section introduced Romoli's paper's results and prerequisite background information. Since the experiments I conduct will replicate Romoli's, it is worth reviewing Romoli's methodology.

The two experiments Romoli conducted share similar procedures. Both experiments asked the participants to read a short description with the format "If A, then B. And A." or "If, then B. But not A". In experiment 1, the participants were asked to select a picture that fit the description the most from four pictures. Each picture can be summarized with three binary categories: whether A is true, whether the presupposition of B is true, and whether B is true. Using this, the four pictures shown are always TTT, TTF, FTF, and FF-.

The descriptions are sorted into two categories – dependent/independent and control/critical. Dependent descriptions exhibit a probable causal relation between A and B, whilst independent descriptions don't. Control descriptions end with "And A.", while critical descriptions end with "But not A". Apart from these, there also exists filler descriptions in which B does not have a relevant presupposition.

Participants were shown 4 control descriptions, 4 critical descriptions, and 8 filler descriptions. The filler and control descriptions were used to verify that the participants do indeed understand the instructions, while the critical descriptions are the actual important measures Romoli wants to measure.

Experiment 2 has the exact same format, though with one important difference – the FTF picture is replaced by a blank picture, and participants were told to choose that if none of the other pictures matched with the description. This is to eliminate the flaw that the FTF picture can match to both types of presuppositions.

3.2 General Method

Translating a person-to-person linguistic experiment into a person-to-LLM experiment in another language is complex. This subsection outlines the general process and the specific LLMs used, while experiment details are covered in their respective sections.

First, all research prompts must be converted into text. While multimodal LLMs exist, they are less common and costly, making text a more practical format. The original experiment must be reviewed to extract essential information. For instance, in Romoli's experiment, the images encode three key pieces of information (see Section 3.1), which must be preserved in text form.

Once all non-text information is transcribed, a prompt mirroring the one given to human subjects is created—usually a simple rewording. The full prompt is provided in Appendix 1.

To ensure methodological rigor, variables should be modified incrementally. Directly translating prompts, images, and subjects into Mandarin for AI processing would introduce multiple confounding variables. Instead, multiple experiments should be conducted, each altering only one key variable at a time.

This study follows a two-step experimental design for each set of experiments. The first replicates Romoli's experiment using English-language LLMs with text-based prompts. The selected models—Gemma 2 9B Instruct, Llama-3.2-3B-Instruct-GGUF, GPT-4o, and Mistral Nemo Instruct—were chosen based on memory efficiency and performance. The second experiment replaces English-language LLMs with Mandarin-based ones, using prompts translated via machine translation and human editing. The Mandarin LLMs tested are glm-4-9b-chat, Spark, Qwen 2.5 Coder 14B, and DeepSeek V2 Lite.

4 Experiments

4.1 Experiment 1a: English LLMs

4.1.1 Setup

As per Romoli, I have written 32 different tests that are of the following format:

(3) Description: If Googlemorph is A, then B. But Googlemorph is not A.
Choice:
A. A creature that is A, p(B) and B.
B. A creature that is A, p(B) and not B.
C. A creature that is not A, is p(B) and not B.
D. A creature that is not A and not p(B).

where p(B) represents the presupposition of B. Within the 32 tests, 16 tests have a causal relationship between A and B (dependent), and 16 tests do not (independent). Apart from this, I have also designed 4 tests with "If Googlemorph is A, then B. And Googlemorph is A." to confirm the logical robustness of the LLMs. The specific 32 prompts will be detailed in Appendix 1 – the following is one example.

(4) Description: If Googlemorph pecks wood, then its beak is sharp. But Googlemorph isn't pecking wood.
Choice:
A. A creature that is pecking wood, has a beak, and the beak is sharp.
B. A creature that is pecking wood, has a beak, and the beak is round.
C. A creature that is not pecking wood, has a beak, and the beak is round.
D. A creature that is not pecking wood and doesn't have a beak.

The tests are conducted with a preamble prompt that details what the LLM agent needs to do – see Appendix 1 for details. All tests are conducted in LMStudio in a Windows 11 environment, and the temperature of all LLMs is set to 0.8, with the exception of ChatGPT 4o, which is conducted on its own website. All analyses were performed using R Statistical Software (R Core Team, 2021).

4.1.2 Rationale For Methodology

This experiment tests the behaviors of the English LLMs in the environment of a presupposition hidden inside a conditional.

I will hereby give an example to demonstrate. Consider example (4). The presupposition "Googlemorph has a beak" is contained within the consequent of the premise "If Googlemorph pecks wood, then its beak is sharp". This can be interpreted in two ways: that there exists a non-conditional presupposition "Googlemorph has a beak", or that there exists a conditional presupposition "If Googlemorph pecks wood, then Googlemorph has a beak".

Now consider choice (D); since in choice (D), the presupposition of "Googlemorph has a beak" is rejected, choice (D) contradicts having an unconditional presupposition. Choice (C) claims the opposite – that "Googlemorph has a beak" is not rejected, and thus is more aligned to the unconditional presupposition.

Of course, choosing choice (C) does not mean that the model isn't forming a conditional presupposition: "If Googlemorph pecks wood, then Googlemorph has a beak. Googlemorph does not peck wood." does not logically link to "Googlemorph doesn't have a beak". This inference, however, would be present in humans because of cognitive bias – specifically, the bias of negating the antecedent. We can, therefore, infer that by choosing (C), the participant is likely to have created a non-conditional presupposition, though we cannot rule out a conditional one.

This conundrum will be rectified in Experiment 2. For now, it would be interesting to see how these factors interact. In order to take this into account, Romoli calculated the percentages of conditional and non-conditional presuppositions based on the assumption that people creating a conditional presupposition will randomly choose between (C) and (D).

4.1.3 Results and Discussion

Table 1 consists of the results, after being converted into percentages and adjust accordingly based on the method in the section above.

Model	C-D	NC-D	C-I	NC-I
Gemma	25%	68.75%	12.5%	85%
Mistral	87.5%	12.5%	87.5%	12.5%
GPT	50%	50%	100%	0%
Llama	0%	81.25%	37.5%	12.5%
Human	71.2%	26.3%	39%	58.2%

Table 1: Experiment 1a Data. C = Conditional, NC = Non-conditional, D = Dependent, I = Independent.

The results are surprising given Romoli's findings. While Gemma shows fewer conditional presuppositions for independent descriptions—matching Romoli—none of the other models do; in fact, two show the opposite pattern. For non-conditional presuppositions, GPT and LLaMa diverge from human behavior, Gemma aligns with it, and Mistral remains unchanged. This therefore suggests that it is the most "humanlike," while GPT and LLaMa diverge the most, and Mistral shows no distinction between dependent and independent descriptions. This is curious, as it means that even though English LLMs trained using human-generated data, they still arrived at a conclusion unlike humans when it comes to presupposition generation, if Romoli's paper is to be considered. The varied behaviors make it difficult to form a unified theory of how English LLMs handle conditional and non-conditional presuppositions, leaving open the question of whether they align with satisfaction theory, DRT, or any single framework.

4.2 Experiment 1b: Mandarin LLMs

4.2.1 Setup

The prompts used in Experiment 1a have been translated through ChatGPT and proofread by a native Mandarin speaker to ensure accuracy. A similar setup is used in Experiment 1a. As mentioned in Section 3.2, the 4 selected LLMs used are glm-4-9b-chat, Spark, Qwen 2.5 Coder 14B, and DeepSeek V2 Lite. Qwen and DeepSeek are tested under the environment given by LMStudio, whilst GLM and Spark are tested using their given APIs in their website.

4.2.2 Results and Discussion

The results are shown in the graph below.

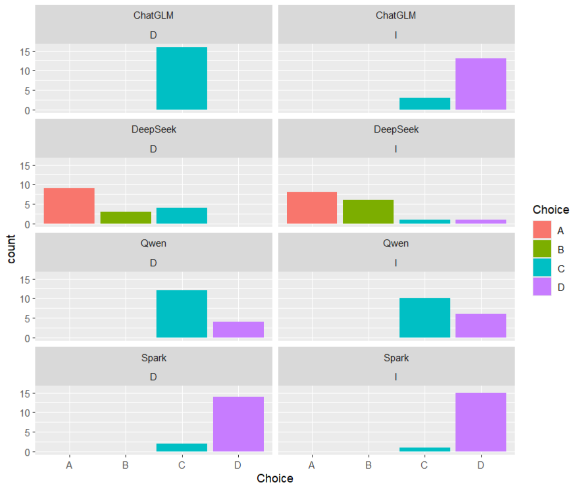

Figure 1: Experiment 1b Model Data

We should be able to employ the same method to transform the data in Experiment 1b to conditional/non-conditional presuppositions. However, this has an issue with any data with a higher count of choice (D) than choice (C). This is because if one considers that there is an equal chance of choosing both (C) and (D) while creating a conditional presupposition, and (C) only when there is a non-conditional presupposition, then (C) should be greater than (D), no matter the exact percentage. However, this is clearly not the case here, as shown in the behaviour of ChatGLM and Spark. Therefore, we can deduce that there must be a preference for choosing (D) when forming conditional presuppositions; this means that the actual conditional presupposition percentage would be smaller than if people were choosing by chance, but bigger than the percentage of choice (D) (as analyzed above, choice (C) includes the possibility of a conditional presupposition). This would not affect the analysis done in Section 4.1.5, apart from introducing uncertainty of the "humanness" of Gemma and the neutrality of Mistral.

To account for this, I will instead use ">" to indicate that the actual percentage is higher than the number shown, and "<" to indicate lower. Thus, we can interpret the data from Section 4.2.2 as follows:

Model	C-D	NC-D	C-I	NC-I
ChatGLM	>0%	<100%	>81.2%	<18.8%
DeepSeek	>0%	<25%	>6.3%	<6.3%
Qwen	>25%	<75%	>37.5%	<62.5%
Spark	>87.5%	<12.5%	>93.8%	<6.3%
Human	71.2%	26.3%	39%	58.2%

Table 2: Experiment 1b Data. C = Conditional, NC = Non-conditional, D = Dependent, I = Independent.

Though we cannot be sure what proportion models are choosing (C) versus (D) when forming a conditional presupposition, they should maintain consistency throughout the experiment; thus, we can use the numbers shown to get an approximate value of the actual proportions of conditional/non-conditional presuppositions.

All four Mandarin LLMs diverged from the human baseline by producing more conditional presuppositions for independent descriptions. This pattern aligns with GPT and LLaMa among the English LLMs, though the degree of difference varies by model. ChatGLM shows the largest percentage shift, but uncertainty prevents precise comparison between conditional and non-conditional counts.

Unfortunately, the uncertainty makes it impossible to provide an exact comparison between conditional and non-conditional presupposition counts. However, by looking at the actual choices, only Qwen matches human behavior by choosing option (C) most often for both dependent and independent descriptions; the other models do not. Overall, Mandarin LLMs consistently produce higher conditional presuppositions for independent descriptions, suggesting a more unified pattern across these models—yet one that still contrasts with human benchmarks.

4.3 Experiment 2a: English LLMs With Covered Box

4.3.1 Setup

The same 32 tests have been modified slightly for the second experiment. The following format is used:

(5) Description: If Googlemorph is A, then B. But Googlemorph is not A.
Choice:
A. A creature that is A, p(B) and B.
B. A creature that is A, p(B) and not B.
C. A creature that is not A and not p(B).
D. None of the above.

As per Romoli, the LLMs are instructed to only select D if there is a better description of the creature than the three choices shown above.

The preamble is detailed in Appendix 1. All tests are conducted in LMStudio in a Windows 11 environment, and the temperature of all LLMs are set to 0.8, with the exception of ChatGPT 4o, which is conducted in its own website.

4.3.2 Rationale For Methodology

The methodology is similar to Experiment 1a and 1b, with the exception of "None of the above", which is the equivalence of the "Covered Box" in Romoli's experiment. Since, as said in Section 4.1.2, rejecting the presupposition in the premise only corresponds to forming a conditional presupposition, replacing the choice of accepting the presupposition to "None of the above" allows the participant to choose said choice only if they have formed a non-conditional presupposition. Thus, by doing so, we can eliminate the effect of accepting the presupposition as an instance of potentially conditional and non-conditional presuppositions.

4.3.3 Results and Discussion

The results are shown in the table below, converted into percentages. The "Human" row refers to Romoli's results; as the covered box solve the issue of converging the two types of presupposition together, we can directly convert responses into percentages.

Model	C-D	NC-D	C-I	NC-I
Gemma	31.25%	68.75%	75%	25%
GPT	100%	0%	100%	0%
LLaMa	37.5%	0%	6.25%	0%
Mistral	18.75%	81.25%	0%	100%
Human	86%	11%	77%	21%

Table 3: Experiment 2a Data. C = Conditional, NC = Non-conditional, D = Dependent, I = Independent.

Experiment 2a showed notable shifts from Experiment 1. Gemma, which previously aligned with human behavior, now increases conditional presuppositions for independent descriptions—opposite of human patterns but consistent with most models in Experiment 1. Mistral now matches human-like variation, reducing conditional presuppositions for independent descriptions and increasing non-conditional ones, a change from its earlier neutrality. Meanwhile, GPT became "neutral," generating only conditional presuppositions regardless of description type. LLaMa similarly changed its non-conditional presuppositions and showed higher error rates when facing independent descriptions.

This drastic behaviour flip between Experiment 1a and 2a could be accounted for by the difference in choices in these two experiments. In Experiment 1a, the choices are explicit; there are no "other possibilities" expressed in them. In Experiment 2a, however, the covered box option is the "catch-all" option that the models can choose if they feel like all three other options do not describe the creature. This may have caused inflation of non-conditional presupposition data, as the models may have thought that none of the three options accurately depict the creature, regardless of whether or not a conditional or a non-conditional presupposition is formed. The drop in conditional presuppositions in Mistral thus may account for this inflation. However, this inflation cannot explain the differences between Gemma, GPT, and LLaMa across experiments.

Section 5 will discuss a proposed solution combining the data of both experiments. Experiment 2b will explore whether this change is consistent cross-linguistically.

4.4 Experiment 2b: Mandarin LLMs With Covered Box

4.4.1 Setup

The prompts used in Experiment 2a have been translated through ChatGPT and proofread by me, a native Mandarin speaker, to ensure accuracy. A similar setup is used in Experiment 2a. As mentioned in Section 3.2, the 4 selected LLMs used are glm-4-9b-chat, Spark, Qwen 2.5 Coder 14B, and DeepSeek V2 Lite. Qwen and DeepSeek are tested under the environment given by LMStudio, whilst GLM and Spark are tested using their given APIs in their website.

4.4.2 Results and Discussion

The results are shown in the table below.

Model	C-D	NC-D	C-I	NC-I
ChatGLM	93.75%	6.25%	6.25%	93.75%
DeepSeek	0%	100%	0%	100%
Qwen	81.25%	18.75%	0%	100%
Spark	0%	50%	0%	75%
Human	86%	11%	77%	21%

Table 4: Experiment 2b Data. C = Conditional, NC = Non-conditional, D = Dependent, I = Independent.

Like English LLMs in Experiment 2a, Mandarin LLMs behave more similarly to human participants from Romoli's experiments. ChatGLM and Qwen, for example, exhibited the behaviour of conditional presuppositions being more prominent in dependent descriptions compared to independent descriptions. Both models also have a higher number of conditional presuppositions than non-conditional presuppositions in dependent descriptions, like that of human behaviour. All models also do not exhibit the opposite behaviour demonstrated in Experiment 1b.

However, there are still some stark differences between human behaviour and Mandarin LLM behaviours. Specifically, there exists a disproportionally high amount of choice (D) in independent descriptions. The "inflation" discussed in Experiment 2a could be the cause of this behaviour. This makes sense, as independent descriptions often include sentences that do not make logical sense, and therefore harder for LLMs to derive more information. This thus caused the LLMs to have to choose (D) much more frequently for independent descriptions due to it being the "catch-all" choice. DeepSeek might have extended this behaviour to even dependent descriptions, thus resulting in the 100% responses choosing (D) in both description types.

This may undermine the percentages of responses signaling a non-conditional presupposition, but I will argue that this would not undermine the overall behaviour of Mandarin LLMs behaving more humanlike in Experiment 2b than in 1b. The reason is that, in the prompt, I have explicitly asked the models to only consider the None of the Above choices if the other three prompts do not fit the description; in other words, the covered box choice is a "last resort". Though there may be responses that ignored this order, the majority of the responses should still follow my prompt. If the covered box choice is considered a "last resort," then models that generate conditional presuppositions will still choose choice (C), as a choice (C) does not run counter to the description in the questions. Thus, though the actual percentage of non-conditional presupposition generation may be lower, it still remains that in Experiment 2b, Mandarin LLMs behave more like humans than in Experiment 1b.

5 Overall Discussion

5.1 Synthesis of Results

Combining the results found in the 4 experiments, one can see that the LLMs have quite a variety of behaviours, whether in Mandarin or English. When an explicit choice of a choice derived from a conditional presupposition and a non-conditional one is given, most LLMs generated conditional presuppositions more when facing independent descriptions – however, when there is only an explicit choice of a conditional presupposition-derived choice and a "None of the Above" choice, most LLMs, especially the Mandarin ones, swapped behaviours, observing a lower percentage of conditional presuppositions in independent descriptions, and correspondingly a higher percentage of non-conditional presuppositions.

This result is quite unusual, as Experiment 1 runs counter-intuitive with what we expected. As said in section 4.1.5, humans tend to expect a conditional presupposition when facing dependent descriptions, unlike what is exhibited here. What is more surprising is that this behaviour is mitigated by simply hiding the option of non-conditional presuppositions in both languages, as shown in Experiment 2. Clearly, the hidden choice created a behaviour change – and though a tentative explanation of "inflation" is discussed in the discussion portions of Experiment 2, it is hard to believe that alone can drastically change the behaviours that much.

A proposed explanation is shown as follows: the key difference between Experiment 1 and 2 is that 1 gives the choice of selection between FF- and FTF – i.e., an explicit selection between conditional and non-conditional presuppositions. Experiment 2, however, does not; one only needs to consider whether the FF- choice, or the conditional presupposition choice, is valid for the description.

In conditional presupposition theory, which DRT is mapped upon, detailed in Section 2.1, conditional presupposition remains the "default choice" in presupposition generation. Only when there is enough justification would a language user choose to defer to a non-conditional presupposition, according to this theory. Here, Occam's Razor and the denying of the antecedent fallacy come into play again – the FF- choice assumes that the presupposition in the consequent of the premise is incorrect, and that is assuming something that cannot be arrived through a conditional presupposition. One cannot arrive from "If Googlemorph can fly, then it has wings" and "Googlemorph cannot fly" to get "It does not have wings" – thus, the covered box choice is selected to avoid the fallacy. In Experiments 1a and 1b, however, the covered box is made explicit – the choice now assumed that the presupposition is true. Thus, since one cannot avoid the fallacy, unless there is other evidence that suggests the presupposition is true – for example, the causality between the antecedent and the consequent of the premise in dependent descriptions – it is best to default to not assuming the truth of the presupposition, hence the increase in FF- choices in independent descriptions in Experiment 1a and 1b.

Since the majority of the tested LLMs' behaviour is straightforwardly explained through conditional presupposition theory, which maps to satisfaction theory, one can thus conclude that the tested LLMs are more likely to subscribe to satisfaction theory, corroborating with the human results of Romoli's paper. However, this does not translate to the idea that **all** LLMs follow the satisfaction theory, especially considering the sheer amount of variation we see in both Experiments.

Why is it that there are so many variations on the presupposition behaviours of LLMs, as we have observed above? The observed variations in presupposition behaviours among LLMs can be attributed to a combination of implicit influences stemming from differences in training data, model architectures, fine-tuning strategies, and linguistic processing mechanisms. While these variations might initially seem unpredictable, they are rooted in systematic factors that govern how each model encodes and interprets linguistic structures.

The most immediate source of variation is the training data itself. Different LLMs are trained on distinct corpora, ranging from structured datasets such as books, academic papers, and formal articles to more unstructured, conversational data such as social media posts and forum discussions. A model that has been exposed to a large amount of structured text is more likely to follow formal linguistic principles, whereas a model trained on noisier, informal data may exhibit less predictable presuppositional behavior.

Architectural differences also play a crucial role. Tokenization strategies, such as Byte-Pair Encoding or SentencePiece, determine how models segment words and phrases, directly influencing how they process presuppositional triggers. Some architectures prioritize syntactic dependencies, allowing them to recognize linguistic structures more explicitly, while others rely more on semantic embeddings, making them more sensitive to meaning rather than rigid grammatical frameworks. Transformer depth and the number of attention layers also impact linguistic reasoning, with larger models generally performing more consistently than smaller ones. However, even among large models, differences in training objectives and internal representations can lead to variations in how presuppositional content is interpreted.

Ultimately, the wide range of presuppositional behaviors observed in LLMs is a reflection of the diversity in training methodologies, architectural design, and linguistic reasoning strategies. While some models demonstrate greater consistency with human-like presupposition projection, others diverge due to differences in how they internalize and retrieve linguistic knowledge. This shows that while LLMs can create convincingly accurate human-like text, this does not mean that they process said text like humans, or even the same between models. This finding is increasingly important as LLM technology becomes more advanced – it is very possible that we would see an increase in the humanness of the generated text, while seeing no change in how LLMs process said text. This disparity between processing and generation should raise alarm and create further research towards resolving said disparity.

5.2 AI As Participants of Linguistic Research

As LLMs become more human-like, it is not a stretch to wonder if they can be analyzed in a human-like way as well—this is one of the motives for the research demonstrated here. I have shown that interesting results can come from using AI as participants in experimental linguistic research, both from verifying the robustness of past linguistic research and from gaining insight into how LLMs work from a linguistic perspective.

Indeed, the usage of LLMs in linguistic research has been explored by other researchers as well. In the paper "Large Language Models and the Wisdom of Small Crowds," Trott demonstrated that LLMs could indeed be useful in linguistic research, specifically as a representation of the aggregate behaviour of many humans ((Trott, 2024)). Though individual variations are less easy to model using LLMs, this lends credibility to the methods used in this paper, as this paper is trying to contribute to the proviso problem debate in a computational manner.

The idea of LLM participants is linked to the "humanness" of LLMs, or how closely they behave like humans. If the humanness of LLM is high, then theoretically, using LLMs in linguistic research would be similar to using humans in linguistic research. Thus, studies like this paper can also act as a benchmark or an evaluation of how much LLMs have evolved. We can see here that Mandarin LLMs are less human-like than English LLMs, given that they exhibit less consistent behaviour than English LLMs, as evidenced by the long explanation given in the previous subsection compared to the relatively simple explanation given for English LLMs. Indeed, the human-likeness benchmark proposed by Duan et al. used 10 psycholinguistic experiments as a basis to assess the humanness of LLMs ((Duan et al., 2024)). Thus, papers like these provide insight not only to linguistics but to the advancement of LLMs as a whole.

6 Conclusion

This dissertation explored the processing of presupposition projection in both English and Mandarin LLMs, drawing comparisons to established linguistic theories – satisfaction theory and Discourse Representation Theory. By adapting experimental methods originally developed for human participants, specifically the paper by Romoli, it was revealed that both English and Mandarin LLMs demonstrate a preference for conditional presuppositions when explicitly tested, aligning them more closely with the Satisfaction Theory framework. However, this generalization is weak, at best, as there exist significant variations among all 8 models tested. This suggests that LLMs can generate convincingly human-like text while lacking the human-like understanding of pragmatic elements, such as pragmatics.

These results contribute to the ongoing debate on LLMs' linguistic capabilities, specifically exploring the potential of using LLMs as participants in experimental linguistic research, both to test theoretical models and to gain insights into the inner workings of AI systems. It is the author's wish that this paper can demonstrate a practical example of LLMs' usefulness in linguistic research.

7 Limitations and Further Research

Because of limitations in computational power and access to LLMs, there are many opportunities for future research following the same vein as this paper. In more detail, the language models used in this paper are not the most powerful, nor do they have the most amount of tokens. As said in Section 3, the choice of language models is also built on whether the experiment can actually be conducted with a Windows 11 system computer with moderate memory and no GPU. Therefore, there is a wide opportunity for new and more powerful LLMs to be tested using this methodology, since they, in theory, would be more in line with human intuition and thus exhibit more human-like behaviours. Specifically, I would like to see how the most recent LLMs, such as GPT o1, Claude Sonnet, and a more powerful version of Qwen, fare with Romoli's experiments.

Moreover, Romoli's paper uses pictures instead of text-based choices. The reason this paper uses text-based choices is the lack of image recognition on several models tested here. It would be interesting to explore how presupposition interacts with multimodal LLMs that can process images so as to more closely replicate Romoli's experiments.

Time constraints also forbid me from conducting multiple trials of a model in testing. Further research can replicate the experiment to verify the results in this paper, either through multiple testing of models using my prompts or alternate prompts from other researchers.

To that end, this paper shows a possible methodology to conduct, evaluate, and possibly improve the underlying linguistic elements of pragmatics in LLMs – not only in English but in other languages as well. I can foresee that this methodology may be used to put LLM participants through already-done psycholinguistic or experimental pragmatics experiments to either verify or investigate said LLMs. This would open up a wide range of further research topics and hopefully help the advancement of not only pragmatics and computational linguistics but also the understanding and high-level explainability of large language models as well.

References

Daiki Asami and Saku Sugawara. 2023. PROPRES: Investigating the Projectivity of Presupposition with Various Triggers and Environments. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 122–137, Singapore. Association for Computational Linguistics.

David I. Beaver, Bart Geurts, and Kristie Denlinger. 2024. Presupposition. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting Language Models for Linguistic Structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Ye-eun Cho and Seongmook Kim. 2024. Pragmatic inference of scalar implicature by LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20.

Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhenguang G. Cai. 2024. Hlb: Benchmarking llms' humanlikeness in language use. *ArXiv*, abs/2409.15890.

Bart Geurts. 1996. Local Satisfaction Guaranteed: A Presupposition Theory and Its Problems. *Linguistics and Philosophy*, 19(3):259–294. Publisher: Springer.

Irene Heim. 2002. On the Projection Problem for Presuppositions. In *Formal Semantics*, pages 249–260. John Wiley & Sons, Ltd.

Xudong Hong, Margarita Ryzhova, Daniel Biondi, and Vera Demberg. 2024. Do large language models and humans have similar behaviours in causal inference with script knowledge? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 421–437, Mexico City, Mexico. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

H. Kamp. 1981. A theory of truth and semantic representation, 277-322, jag groenendijk, tmv janssen and mbj stokhof, eds. In Jeroen A. G. Groenendijk, editor, *Formal methods in the study of language*. U of Amsterdam.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A Corpus of Naturally-Occurring Presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Jacopo Romoli. 2011. An experimental investigation of presupposition projection in conditional sentences. *Semantics and Linguistic Theory*, 21:592–608.

Judith Sieker and Sina Zarrieß. 2023. When Your Language Model Cannot Even Do Determiners Right: Probing for Anti-Presuppositions and the Maximize Presupposition! Principle. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting

Neural Networks for NLP, pages 180–198, Singapore. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12075–12097, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sean Trott. 2024. Large language models and the wisdom of small crowds. *Open Mind*, 8:723–738.

Polina Tsvilodub, Paul Marty, Sonia Ramotowska, Jacopo Romoli, and Michael Franke. 2024. Experimental Pragmatics with Machines: Testing LLM Predictions for the Inferences of Plain and Embedded Disjunctions. *arXiv preprint*. ArXiv:2405.05776 [cs].

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

A Prompts

A.1 Experiment 1a Preamble

I will present to you two pieces of data; the first one is labeled "Description" and the second one is labeled "Choice". The description describes one of the creatures shown in the possible choices. Please read the following description. Depending on the description, select the best creature that fits this description. You do not need justification.

The following snippets are examples. Description: If Googlemorph is diving, then his scales are sleek. But Googlemorph is not diving.

Choice:

A. A creature that is diving, with scales, and the scales are sleek.

B. A creature that is diving, with scales, and the scales are rough.

C. A creature that is on the ground, with scales, and the scales are rough.

D. A creature that is on the ground, without scales.

Answer: C

Description: If Googlemorph is blue, then his scales are sleek. But Googlemorph is not diving.

Choice:

A. A creature that is blue, with scales, and the scales are sleek.

B. A creature that is blue, with scales, and the scales are rough.

C. A creature that is green, without scales.

D. A creature that is green, with scales, and the scales are rough.

Answer: D

A.2 Experiment 1a Prompts

1. Description: If Googlemorph is flying, then his wings are big and strong. But Googlemorph is not flying.

   Choice:

   A. A creature that is flying, with wings, and the wings are big.

   B. A creature that is flying, with wings, and the wings are small.

   C. A creature that is on the ground, with wings, and the wings are small.

   D. A creature that is on the ground, without wings.

2. Description: If Googlemorph is drinking orange juice, then his wings are big and strong. But Googlemorph is not drinking orange juice.

   Choice:

   A. A creature that is drinking orange juice, with wings, and the wings are big.

   B. A creature that is drinking orange juice, with wings, and the wings are small.

   C. A creature that is drinking water, with wings, and the wings are small.

   D. A creature that is drinking water, without wings.

3. Description: If Googlemorph has sharp teeth, then he is eating meat. But Googlemorph doesn't have sharp teeth.

   Choice:

   A. A creature that has sharp teeth, is eating, and the food is meat.

   B. A creature that has sharp teeth, is eating, and the food is plants.

   C. A creature that doesn't have sharp teeth, is eating, and the food is plants.

   D. A creature that doesn't have sharp teeth, and is not eating.

4. Description: If Googlemorph is green, then he is eating meat. But Googlemorph is not green.

   Choice:

   A. A creature that is green, is eating, and the food is meat.

   B. A creature that is green, is eating, and the food is plants.

   C. A creature that is blue, is eating, and the food is plants.

   D. A creature that is blue, and is not eating.

5. Description: If Googlemorph is underwater, then its gills are functioning. But Googlemorph is not underwater.

   Choice:

   A. A creature that is underwater, has gills, and the gills are functioning.

   B. A creature that is underwater, has gills, and the gills are not functioning.

   C. A creature that is on the ground, has gills, and the gills are not functioning.

   D. A creature that is on the ground and doesn't have gills.

6. Description: If Googlemorph is green, then its gills are functioning. But Googlemorph is not green.

   Choice:

   A. A creature that is green, has gills, and the gills are functioning.

   B. A creature that is green, has gills, and the gills are not functioning.

   C. A creature that is blue, has gills, and the gills are not functioning.

   D. A creature that is blue and doesn't have gills.

7. Description: If Googlemorph breathes fire, then its snout is fireproof. But Googlemorph isn't breathing fire.

   Choice:

   A. A creature that breathes fire, has a snout, and the snout is fireproof.

   B. A creature that breathes fire, has a snout, and the snout is not fireproof.

   C. A creature that doesn't breathe fire, has a snout, and the snout is not fireproof.

   D. A creature that doesn't breathe fire, and doesn't have a snout.

8. Description: If Googlemorph is pink, then its snout is fireproof. But Googlemorph isn't pink.

   Choice:

   A. A creature that is pink, has a snout, and the snout is fireproof.

   B. A creature that is pink, has a snout, and the snout is not fireproof.

   C. A creature that is purple, has a snout, and the snout is not fireproof.

   D. A creature that is purple, and doesn't have a snout.

9. Description: If Googlemorph runs, then its legs are strong. But Googlemorph isn't running.

   Choice:

   A. A creature that is running, has legs, and the legs are strong.

   B. A creature that is running, has legs, and the legs are not strong.

   C. A creature that is still, has legs, and the legs are not strong.

   D. A creature that is still and doesn't have legs.

10. Description: If Googlemorph is orange, then its legs are strong. But Googlemorph isn't orange.

    Choice:

    A. A creature that is orange, has legs, and the legs are strong.

    B. A creature that is orange, has legs, and the legs are not strong.

    C. A creature that is red, has legs, and the legs are not strong.

    D. A creature that is red and doesn't have legs.

11. Description: If Googlemorph sees far, then its eyes are large. But Googlemorph isn't seeing far.

    Choice:

A. A creature that sees far, has eyes, and the eyes are large.

B. A creature that sees far, has eyes, and the eyes are not large.

C. A creature that doesn't see far, has eyes, and the eyes are not large.

D. A creature that doesn't see far and doesn't have eyes.

12. Description: If Googlemorph is blue, then its eyes are large. But Googlemorph isn't blue.

    Choice:

    A. A creature that is blue, has eyes, and the eyes are large.

    B. A creature that is blue, has eyes, and the eyes are not large.

    C. A creature that is yellow, has eyes, and the eyes are not large.

    D. A creature that is yellow and doesn't have eyes.

13. Description: If Googlemorph uses a slingshot, then its hands have thumbs. But Googlemorph isn't using a slingshot.

    Choice:

    A. A creature that uses a slingshot, has hands, and the hands have thumbs.

    B. A creature that uses a slingshot, has hands, and the hands don't have thumbs.

    C. A creature that doesn't use a slingshot. has hands, and the hands don't have thumbs.

    D. A creature that doesn't use a slingshot and doesn't have hands.

14. Description: If Googlemorph is red, then its hands have thumbs. But Googlemorph isn't red.

    Choice:

    A. A creature that is red, has hands, and the hands have thumbs.

    B. A creature that is red, has hands, and the hands don't have thumbs.

    C. A creature that is green. has hands, and the hands don't have thumbs.

    D. A creature that is green and doesn't have hands.

15. Description: If Googlemorph has a good memory, then its hippocampus is developed. But Googlemorph doesn't have a good memory.

    Choice:

    A. A creature that has a good memory, has a hippocampus, and the hippocampus is developed.

    B. A creature that has a good memory, has a hippocampus, and the hippocampus is not developed.

    C. A creature that doesn't have a good memory, has a hippocampus, and the hippocampus is developed.

    D. A creature that doesn't have a good memory and doesn't have a hippocampus.

16. Description: If Googlemorph is orange, then its hippocampus is developed. But Googlemorph isn't orange.

    Choice:

    A. A creature that is orange, has a hippocampus, and the hippocampus is developed.

    B. A creature that is orange, has a hippocampus, and the hippocampus is not developed.

    C. A creature that is blue, has a hippocampus, and the hippocampus is developed.

    D. A creature that is blue and doesn't have a hippocampus.

17. Description: If Googlemorph sees in color, then its eyes contain three types of cones. But Googlemorph doesn't see in color.

    Choice: A. A creature that sees in color, has eyes, and the eyes contain three types of cones.

    B. A creature that sees in color, has eyes, and the eyes don't contain three types of cones.

    C. A creature that doesn't see in color, has eyes, and the eyes don't contain three types of cones.

    D. A creature that doesn't see in color and doesn't have eyes.

18. Description: If Googlemorph is white, then its eyes contain three types of cones. But Googlemorph isn't white.

    Choice:

A. A creature that is white, has eyes, and the eyes contain three types of cones.

B. A creature that is white, has eyes, and the eyes don't contain three types of cones.

C. A creature that is black, has eyes, and the eyes don't contain three types of cones.

D. A creature that is black and doesn't have eyes.

19. Description: If Googlemorph communicates, then its Broca's area is functioning. But Googlemorph doesn't communicate.

    A. A creature that is communicating, has a Broca's area, and the Broca's area is functioning.

    B. A creature that is communicating, has a Broca's area, and the Broca's area is not functioning.

    C. A creature that is not communicating, has a Broca's area, and the Broca's area is not functioning.

    D. A creature that is not communicating and doesn't have a Broca's area.

20. Description: If Googlemorph is purple, then its Broca's area is functioning. But Googlemorph isn't purple.

    A. A creature that is purple, has a Broca's area, and the Broca's area is functioning.

    B. A creature that is purple, has a Broca's area, and the Broca's area is not functioning.

    C. A creature that is gray, has a Broca's area, and the Broca's area is not functioning.

    D. A creature that is gray and doesn't have a Broca's area.

21. Description: If Googlemorph eats chocolate, then its liver can process theobromine efficiently. But Googlemorph doesn't eat chocolate.

    A. A creature that is eating chocolate, has a liver, and the liver process theobromine efficiently.

    B. A creature that is eating chocolate, has a liver, and the liver doesn't process theobromine efficiently.

    C. A creature that is not eating chocolate, has a liver, and the liver doesn't process theobromine efficiently.

    D. A creature that is not eating chocolate and doesn't have a liver.

22. Description: If Googlemorph has 37 teeth, then its liver can process theobromine efficiently. But Googlemorph doesn't have 37 teeth.

    A. A creature that has 37 teeth, has a liver, and the liver process theobromine efficiently.

    B. A creature that has 37 teeth, has a liver, and the liver doesn't process theobromine efficiently.

    C. A creature that has 31 teeth, has a liver, and the liver doesn't process theobromine efficiently.

    D. A creature that has 31 teeth and doesn't have a liver.

23. Description: If Googlemorph pecks wood, then its beak is sharp. But Googlemorph isn't pecking wood.

    A. A creature that is pecking wood, has a beak, and the beak is sharp.

    B. A creature that is pecking wood, has a beak, and the beak is round.

    C. A creature that is not pecking wood, has a beak, and the beak is round.

    D. A creature that is not pecking wood and doesn't have a beak.

24. Description: If Googlemorph is gray, then its beak is sharp. But Googlemorph isn't gray.

    A. A creature that is gray, has a beak, and the beak is sharp.

    B. A creature that is gray, has a beak, and the beak is round.

    C. A creature that is purple, has a beak, and the beak is round.

    D. A creature that is purple and doesn't have a beak.

25. Description: If Googlemorph hears a whale sing, then its ears can hear subsonic sounds. But Googlemorph doesn't hear a whale sing.

    A. A creature that hears a whale sing, has ears, and the ears can hear subsonic sounds.

    B. A creature that hears a whale sing, has ears, and the ears cannot hear subsonic sounds.

C. A creature that doesn't hear a whale sing, has ears, and the ears cannot hear subsonic sounds.

D. A creature that doesn't hear a whale sing and doesn't have ears.

26. Description: If Googlemorph is blue, then its ears can hear subsonic sounds. But Googlemorph isn't blue.

    A. A creature that is blue, has ears, and the ears can hear subsonic sounds. B. A creature that is blue, has ears, and the ears cannot hear subsonic sounds. C. A creature that is yellow, has ears, and the ears cannot hear subsonic sounds. D. A creature that is yellow and doesn't have ears.

27. Description: If Googlemorph lives in the Arctic, then its fur is thick. But Googlemorph doesn't live in the Arctic.

    A. A creature that lives in the Arctic, has fur, and the fur is thick.

    B. A creature that lives in the Arctic, has fur, and the fur is thin.

    C. A creature that lives in the Equator, has fur, and the fur is thin.

    D. A creature that lives in the Equator and doesn't have fur.

28. Description: If Googlemorph smiles, then its fur is thick. But Googlemorph isn't smiling.

    A. A creature that is smiling, has fur, and the fur is thick.

    B. A creature that is smiling, has fur, and the fur is thin.

    C. A creature that is crying, has fur, and the fur is thin.

    D. A creature that is crying and doesn't have fur.

29. Description: If Googlemorph eats plants, then its teeth is flat. But Googlemorph isn't eating plants.

    Choice:

    A. A creature that is eating plants, has teeth, and the teeth is flat.

    B. A creature that is eating plants, has teeth, and the teeth is sharp.

    C. A creature that is not eating plants, has teeth, and the teeth is sharp.

    D. A creature that is not eating plants and doesn't have teeth.

30. Description: If Googlemorph is pink, then its teeth is flat. But Googlemorph isn't pink.

    Choice:

    A. A creature that is pink, has teeth, and the teeth is flat.

    B. A creature that is pink, has teeth, and the teeth is sharp.

    C. A creature that is purple, has teeth, and the teeth is sharp.

    D. A creature that is purple and doesn't have teeth.

31. Description: If Googlemorph jumps high, then its legs are strong. But Googlemorph doesn't jump high.

    Choice:

    A. A creature that is jumping high, has legs, and the legs are strong.

    B. A creature that is jumping high, has legs, and the legs are weak.

    C. A creature that is on the ground, has legs, and the legs are weak.

    D. A creature that is on the ground and has no legs.

32. Description: If Googlemorph is blue, then its legs are strong. But Googlemorph isn't blue.

    Choice:

    A. A creature that is blue, has legs, and the legs are strong.

    B. A creature that is blue, has legs, and the legs are weak.

    C. A creature that is pink, has legs, and the legs are weak.

    D. A creature that is pink and has no legs.

A.3 Experiment 2a Preamble

I will present to you two pieces of data; the first one is labeled "Description" and the second one is labeled "Choice". The description describes one of the creatures shown in the possible choices. Please read the following description. Depending on the description, select the best creature that fits this description. You do not need justification.

Note: The prompt in Experiment 2a is modified per Section 4.3.1 based on Appendix A.1's prompts.